Skew Brownian Motion and Complexity of the ALPS Algorithm

Gareth O. Roberts[1], Jeffrey S. Rosenthal[2], and Nicholas G. Tawn[3]

[1]*Department of Statistics, University of Warwick, United Kingdom, CV4 7AL,* Gareth.O.Roberts@warwick.ac.uk

[2]*Department of Statistical Sciences, University of Toronto, 100 St. George Street, Room 6018, Toronto, Ontario, Canada M5S 3G3,* jeff@math.toronto.edu

[3]*Department of Statistics, University of Warwick, United Kingdom, CV4 7AL,* n.tawn.1@warwick.ac.uk

(September 2020; last revised April 2021)

Abstract

Simulated tempering is a popular method of allowing MCMC algorithms to move between modes of a multimodal target density π. The paper [29] introduced the Annealed Leap-Point Sampler (ALPS) to allow for rapid movement between modes. In this paper, we prove that, under appropriate assumptions, a suitably scaled version of the ALPS algorithm converges weakly to skew Brownian motion. Our results show that under appropriate assumptions, the ALPS algorithm mixes in time $O(d[\log d]^2)$ or $O(d)$, depending on which version is used.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms [6] are very widely used to explore and sample from a complicated high-dimensional target probability distribution π. The most basic version of MCMC is the *Metropolis algorithm* [17, 10]. From a given state x, it proceeds by first *proposing* to move to a new state y, and then either *accepting* that proposal (i.e., moving to y), or *rejecting* that proposal (i.e., staying at x). The *acceptance probability* is given by $\min[1, \pi(y) / \pi(x)]$. If the proposal densities are symmetric (i.e., have the same probability of proposing y from x, as of proposing x from y), this procedure ensures that the resulting Markov chain will be reversible with respect to π, and thus have π as its stationary density.

MCMC algorithms have a tendency to get stuck in local modes, which limits their effectiveness. Annealing and tempering methods [18, 12, 1, 9, 16] attempt to overcome this problem by considering different powers π^β of the target density, where $\beta \le 1$ is an *inverse-temperature*. Here $\beta = 1$ corresponds to the desired distribution, so those are the only

samples which are "counted". However, small positive values $\beta \ll 1$ make the density flatter and thus much easier to traverse.

Despite the tremendous success of tempering, these methods suffer from deficiencies, especially in high dimensions. In particular, tempering of distributions does not usually preserve the relative mass contained in each of the modes. To deal with this, the paper [31] introduced a *weight-preserving transformation* which overcomes the weight instability problem as long as all modes look reasonably Gaussian. Unfortunately, in applications that is often not the case, since modes often exhibit significant skewness.

An alternative approach, the Annealed Leap-Point Sampler (ALPS), was introduced in [29]. This algorithm instead considers very *large* values $\beta \gg 1$, corresponding to very peaked target densities at very cold temperatures. (Large β are often used in optimisation algorithms such as *simulated annealing* [18, 12, 1], but are not normally used by sampling algorithms.) Assuming smoothness, the resulting sharply peaked modes then become approximately Gaussian, thus facilitating simpler ways of moving between them. Furthermore, a weight-preserving transformation is performed to approximately preserve the probabilistic weight of each peak upon tempering.

For any MCMC algorithm, an important question is how quickly it converges to its stationary distribution π. While there have been many attempts to bound MCMC convergence times directly (see e.g. [26] and the references therein), much of the effort has been focused on questions of *computational complexity*, i.e. how the algorithm's running time grows as a function of other parameters (dimension, size of data, etc.).

One promising, though technically challenging, approach to determining the computational complexity of Metropolis algorithms is through the use of *diffusion limits* as the dimension $d \to \infty$. Similar to how symmetric random walk converges to Brownian motion under appropriate rescaling, certain transformations of some Metropolis algorithm components will converge to Langevin diffusions. This was originally exploited in [21, 22] to derive complexity and optimality results for ordinary random-walk-based Metropolis algorithms, and was later generalised to many other contexts [23, 4, 25]. Furthermore, the $d \to \infty$ limit of MCMC algorithms also provides good approximate information about processes of modest finite dimension; see e.g. [23, Figure 4].

In this paper, will shall apply the diffusion limits methodology to a "vanilla" version of the ALPS algorithm, to study its convergence complexity. We will prove (Theorem 4) that, under appropriate assumptions, a suitably scaled version of this ALPS algorithm converges to *skew Brownian motion* (cf. [14]). This limit will allow us to draw conclusions about the computational complexity of our algorithm, and to show (Corollaries 5 and 6) that under appropriate assumptions, as the dimension $d \to \infty$ the vanilla ALPS algorithm mixes in time $O(d[\log d]^2)$ or $O(d)$ depending on which version is used.

These results show that ALPS converges fairly quickly even in high dimension. This complexity order is similar to those previously derived for ordinary random-walk Metropolis [21] and for Simulated Tempering [25], which were each shown to converge to dimension-free diffusions when sped up by a factor of d, thus showing that their complexity is $O(d)$. The difference is that those previous results assumed an iid target of the form (1) and (3) with $J = 1$, and assumed immediate mixing between all modes at each β, so it omitted the issue of moving between modes which often makes those algorithms exponentially slow [32]. By

contrast, the ALPS algorithm stores mode location information that is used in a special mode-jumping move to converge efficiently even when there are $J > 1$ widely-separated modes, as we now describe.

2 The ALPS Algorithm

Tempering methods for MCMC usually consider powers π^β for *small* values $\beta \in (0, 1]$, to make the target distribution flatter and thus allow for easier mixing between modes. By contrast, the paper [29] introduced the *Annealed Leap-Point Sampler* (ALPS) algorithm, which instead uses *large* values $\beta \gg 1$, combined with locally weight-preserving tempering distributions as in [31] so the modes retain their relative masses. These choices make the modes of π even more separated. However, under certain smoothness and integrability assumptions, they also make each mode appear approximately Gaussian and hence similarly-shaped. This allows for auxiliary "mode-jumping" Markov chain steps which move effectively between the different modes when β is large. Then, as usual, only samples in the original temperature $\beta = 1$ are "counted" as actual samples from π.

To illustrate the idea of this algorithm, consider the following simple example in dimension $d = 5$. Suppose the target density π on $\mathbf{R}^5$ is a mixture of two skew-normal modes centered at $(-20, -20, -20, -20, -20)$ and $(20, 20, 20, 20, 20)$ respectively, with scalings 1 and 2 respectively, and with shape parameter $\alpha = 10$, so for all $\theta \in \mathbf{R}^5$,

$$\pi(\theta) \;=\; (0.7) \prod_{i=1}^{5} 2\,\phi\big(\theta_i + 20\big)\,\Phi\big(10(\theta_i + 20)\big) \;+\; (0.3) \prod_{i=1}^{5} \phi\big(\frac{1}{2}(\theta_i - 20)\big)\,\Phi\big(5(\theta_i - 20)\big)\,,$$

where as usual $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(u)\,du$; see Figure 1.

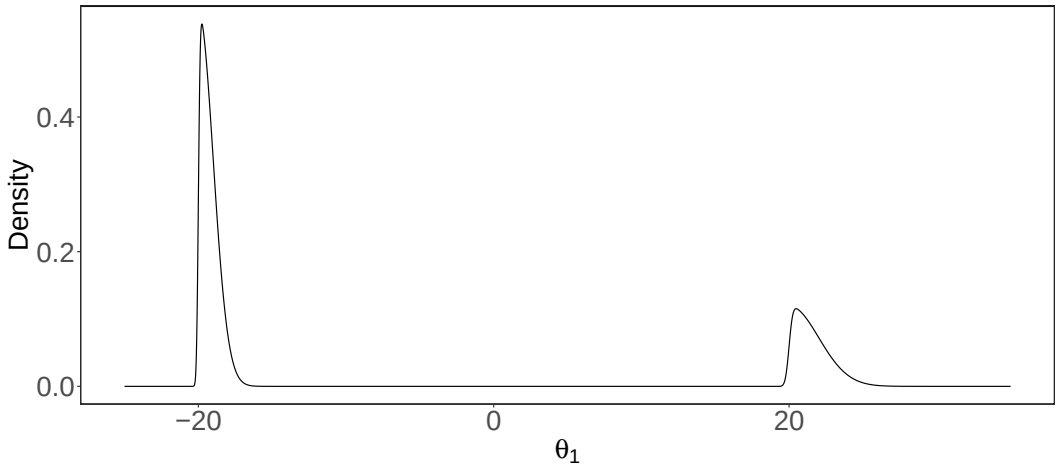

Figure 1: The θ_1 marginal of the target density in the illustrative example.

In such an example, it is very easy for a Markov chain to mix separately *within* either of the two modes. The challenge is to move between the modes (which is virtually impossible for a

typical fixed-temperature Metropolis algorithm even in this simple 5-dimensional example). The ALPS algorithm introduces a powerful independence-sampler-based move so that at very large inverse-temperature values $\beta \gg 1$, the chain can exploit the near-Gaussianity of each of the modes to directly jump between them. Figure 2 shows a trace plot of the inverse-temperature values β during one run of the algorithm, and also indicates by colour which of the two modes the chain is in (i.e., closest to). As can be seen from the plot, the chain stays in the same mode for long periods of time, and only switches modes when the values of β are very large at which point it jumps to either mode with its correct probability. (Note that this description is for the "vanilla" version of ALPS; see Remark 1 below.)

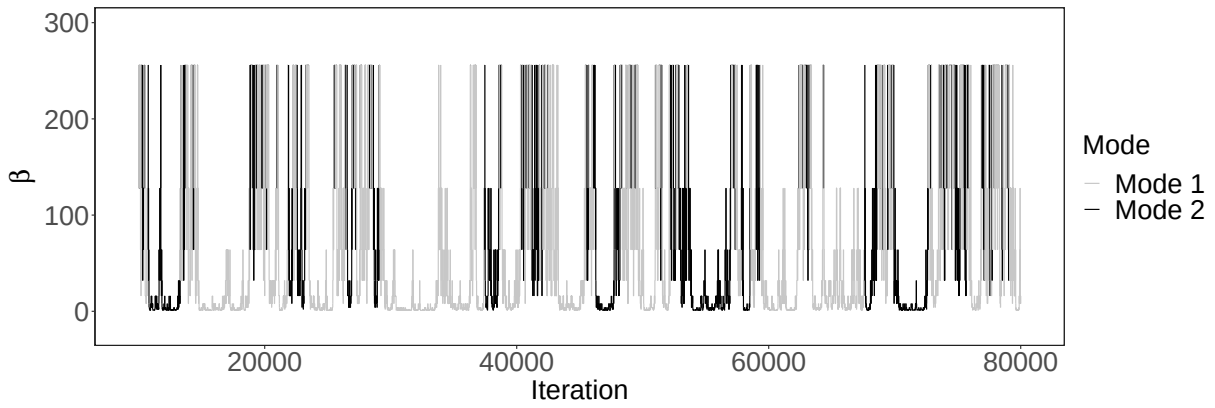

Figure 2: Trace plot of the β values in the illustrative example, coloured to indicate whether the chain is in mode 1 (light-gray) or mode 2 (black).

Figure 2 illustrates that the key to the ALPS algorithm's success is moving rapidly between the large $\beta = \beta_{max} = 256$ values (which allow for mixing between the modes) and the small $\beta = 1$ value (which can be "counted" as a sample from π). However, it is not clear how quickly such mixing takes place, and in particular how it changes depending on the target π and dimension d. To study this, we would like to prove a diffusion limit of a suitably scaled version of the β process, but it is not clear from Figure 2 what sort of limiting diffusive behaviour is available.

To better understand this algorithm's convergence, we consider a suitable transformation of β. Namely, we instead consider the values of $s \log(\beta_{max}/\beta)$, where $s = 1$ if the chain is in mode 1 or $s = -1$ if the chain is in mode 2. The resulting process is shown in Figure 3, which suggests that this modified functional does indeed start to resemble a diffusive process. Indeed, away from the special value 0 (corresponding to β_{max} and the mode-jumping moves), the process looks roughly like Brownian motion. In fact, we shall prove below (Theorem 4) that under appropriate assumptions and scalings, this modified process converges to a skew Brownian motion.

More precisely, we shall prove diffusion limits for suitably rescaled versions of the ALPS algorithm, as the dimension $d \to \infty$. We shall assume that the ALPS algorithm can easily jump between modes when it reaches the sufficiently large inverse-temperature $\beta_{max}^{(d)}$, but that it is stuck within one mode whenever $\beta < \beta_{max}^{(d)}$. We therefore focus on how the inverse-temperatures β themselves are updated by the algorithm. In particular, we will

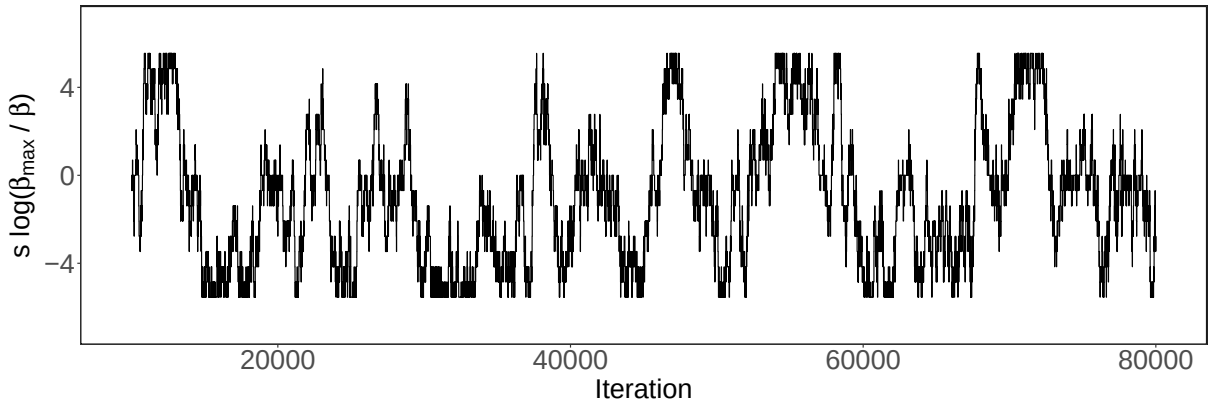

Figure 3: A trace plot of the transformed values $s \log(\beta_{max}/\beta)$ in the illustrative example, where $s = +1$ or $s = -1$ when the chain is in mode 1 or 2.

prove (Theorem 4) that a particular rescaling of the β process converges to *skew Brownian motion* [14]. This will in turn allow us to derive computational complexity results (Section 5).

Remark 1 The "vanilla" ALPS algorithm studied herein differs in certain ways from the full ALPS algorithm for actual applications in [29]. For example, we assume the process mixes perfectly between modes when $\beta = \beta_{max}^{(d)}$ (due to near-Gaussianity and the algorithm's auxiliary mode-jumping steps), and not at all when $\beta < \beta_{max}^{(d)}$, while the full algorithm mixes better and better at higher β values but never perfectly. Also, the full algorithm actually uses *parallel tempering*, in which a separate chain is run at each temperature and their values are swapped; the single β process studied herein can then be thought of as following which of the chains is currently carrying state information between larger and smaller inverse-temperatures and thus facilitating mixing (cf. Section 4 of [2]). Finally, the full ALPS algorithm in [29] also makes use of the QuanTA transformation [30], an additional affine transformation to increase the efficiency of the temperature-swap moves, which we omit here; we discuss the effect of this extra QuanTA transformation in Corollary 6 below.

3 Assumptions

We consider a version of the Annealed Leap-Point Sampler (ALPS) algorithm of [29]. We assume the chain always mixes immediately within each mode, but the chain can only jump between modes when at the sufficiently cold inverse-temperature $\beta = \beta_{max}^{(d)}$, at which point it immediately jumps to any of its modes with the correct probability weight.

To facilitate theoretical analysis, we assume that the target density π is a mixture of J normalised densities $g_1, \ldots, g_J$ on $\mathbf{R}^d$ with weights $w_1, \ldots, w_J$, i.e.

$$\pi(x) \;=\; \sum_{j=1}^{J} w_j g_j(x)\,, \quad x \in \mathbf{R}^d\,. \tag{1}$$

We do not require the g_j to be unimodal, but we shall nevertheless refer to them informally as the "modal components" or "modes" of π, with the intuition that it is easy for MCMC to mix efficiently within each individual g_j but difficult for it to jump between the different g_j.

We also assume that each state x is "allocated" to (i.e. is "in") one of the modes (e.g. whichever one's center it is closest to), such that the accept/reject probabilities when updating β can be computed using only the mode g_j of the current state, rather than the full density π. (This corresponds to considering the x values as elements of $(\mathbf{R}^d)^J$, with a different version of the state space $\mathbf{R}^d$ for each of the J modes; if the modes are well separated, then especially for large β this will be a good approximation to the actual algorithm.)

Then, for each inverse-temperature $\beta \geq 1$, we shall use the tempered distribution

$$\pi_\beta(x) \;\propto\; \sum_{j=1}^{J} w_j \; \frac{[g_j(x)]^\beta}{\int [g_j(x)]^\beta dx} \;=:\; \sum_{j=1}^{J} w_j \, g_j^\beta(x)\,, \tag{2}$$

where g_j^β are the normalised powers of the g_j. We assume the same weights w_j can be used for each β due to a weight-preserving transformation as in [31].

In terms of these assumptions, the vanilla ALPS algorithm as we shall study it is defined by Algorithm 1.

Require: A mixture target distribution π on $\mathbf{R}^d$ as in (1).
Require: A sequence of inverse-temperatures $1 = \beta_0 < \beta_1 < \ldots < \beta_k =: \beta_{max}$.
Require: An initial state $X_0 \in \mathcal{X}$ and inverse-temperature $\beta(0) := \beta_{I(0)}$.
  for $n = 0, 1, 2, 3, \ldots$ **do**
    # State-Changing Phase:
    if $\beta(n+1) = \beta_{max}$ **then**
      Sample $\bar{j} \in \{1, 2, \ldots, J\}$ with probabilities w_j. **(auxiliary mode-jumping)**
    else
      Let $\bar{j}$ denote the mode that the current state X_n is in.
    end if
    Sample $X_{n+1} \sim g_{\bar{j}}$. **(Mix immediately within the current mode only.)**
    # Temperature-Changing Phase:
    Select a proposed new inverse-temperature $\beta_{I_*} = \beta_{I(n)\pm 1}$ with probability 1/2 each.
    Set $\alpha \leftarrow \min\left[1, \frac{g_{\bar{j}}^{\beta_{I_*}}(X_n)}{g_{\bar{j}}^{\beta_{I(n)}}(X_n)}\right]$, with $g_{\bar{j}}^\beta$ as in (2). (Take $\alpha = 0$ if $I_* = 0$ or $I_* = k+1$).
    With probability α, accept the proposal by setting $I(n+1) = I_*$,
      Else, reject the proposal by setting $I(n+1) = I(n)$.
    Set $\beta(n+1) \leftarrow \beta_{I(n+1)}$.
  end for

Algorithm 1: The Vanilla ALPS Algorithm

In our theoretical proofs below, we assume for simplicity (though see Remark 3 below) that we have just $J = 2$ modes, of weights w_1 and $w_2 = 1 - w_1$ respectively. To achieve limiting diffusions, we further assume as in the original MCMC diffusion limit results [21]

that each of the individual components g_j consists of iid univariate coordinates, i.e. that for each j we have

$$g_j(x) \;=\; \prod_{i=1}^{d} \overline{g}_j(x_i) \tag{3}$$

for some fixed one-dimensional density function $\overline{g}_j$, where $x = (x_1, x_2, \ldots, x_d)$. This allows us to apply the diffusion-limit results of [25] within each individual target mode. (Although (3) is a very restrictive assumption, it is known [23] that conclusions drawn from this special case are often approximately applicable in much broader contexts.)

We also require assumptions on the β values. Write the inverse-temperatures as

$$1 = \beta_0^{(d)} < \beta_1^{(d)} < \ldots < \beta_{k(d)}^{(d)} =: \beta_{max}^{(d)} \tag{4}$$

for the process in dimension d. Similar to [2] and [25], following [19] and [13], we assume that the inverse temperatures are related by

$$\beta_i \;=\; \beta_{i-1} + \ell(\beta_{i-1})/d^{1/2} \tag{5}$$

for some fixed C^1 function ℓ. It is shown in [2, 25] that in the single-mode iid case, the fastest limiting diffusion is obtained by using the choice

$$\ell(\beta) \;=\; I^{-1/2}(\beta)\, \ell_0 \tag{6}$$

for a fixed constant $\ell_0 \doteq 2.38$, where $I(\beta) = \mathrm{Var}_{x\sim\overline{g}^\beta}(\log \overline{g}(x))$. Inspired by this, in our later results we shall assume the *Proportionality Condition* that the quantities $I_j(\beta) := \mathrm{Var}_{x\sim\overline{g}_j^\beta}(\log \overline{g}_j(x))$ for the different modes are proportional, i.e. there are positive constants r_j and a C^1 function $I_0 : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$I_j(\beta) \;=\; I_0(\beta)/r_j\,, \qquad j = 1, \ldots, J\,, \tag{7}$$

and shall then correspondingly assume that

$$\ell(\beta) \;=\; I_0^{-1/2}(\beta)\, \ell_0\,, \tag{8}$$

for some fixed constant $\ell_0 > 0$.

One example is the *Exponential Power Family* case, in which each of the mixture component factors g_j is of the form $g_j(x) \propto e^{-\lambda_j |x|^{r_j}}$ for some $\lambda_j, r_j > 0$. It then follows from Section 2.4 of [2] that $I_j(\beta) = \beta^{-2}/r_j$ for $\beta > 0$, so the Proportionality Condition (7) holds with $I_0(\beta) = \beta^{-2}$. The corresponding choice of ℓ from (6) is then $\ell(\beta) = \beta/\sqrt{r_j}$ in mode j. This includes the Gaussian case, where each $r_j = 2$ and $\lambda_j = 1/\sigma_j^2$.

Remark 2 Our assumptions of immediate mixing within modes, and immediate mixing between modes when $\beta = \beta_{max}$, is analogous to the corresponding assumptions in [31, Section 5.1] for ordinary Simulated Tempering of immediate mixing within modes, and immediate mixing between modes when $\beta = \beta_{min}$ (the hottest temperature). In practice, even within simple modes the mixing is not immediate, but rather takes e.g. $O(d)$ iterations for

random-walk Metropolis (RWM) [21], or $O(d^{1/3})$ for Langevin algorithms [22], or $O(d^{1/4})$ for Hamiltonian (Hybrid) Monte Carlo [5]. Since we shall show that the β-mixing for ALPS takes at least $O(d)$, it follows that if Langevin or Hamiltonian dynamics are used for the state-changing phase, then the states will mix at a faster order than the temperatures, thus effectively immediately, in which case our assumption of immediate mixing within modes is reasonable. By contrast, if RWM dynamics are used for the state-changing phase, then the interplay between the temperature convergence and state convergence would be more complicated, though it still works effectively in practice [29].

4 Main Results

We now state various weak convergence results for various transformations of our process. (All proofs are deferred to Section 6 below.) Let $\beta^{(d)}(t)$ be the inverse temperature at time t for the process described by Algorithm 1 in dimension d. Let $\beta^{(d)}(N(dt))$ be a continuous-time version of the $\beta^{(d)}(t)$ process, sped up by a factor of d, where $\{N(t)\}$ is an independent standard rate-1 Poisson process. To combine the two modes into one single process, we further augment this process by multiplying it by -1 when the algorithm's state is allocated to the second mode, while leaving it positive (unchanged) when state is allocated to the first mode, i.e.

$$X_t^{(d)} \;=\; \begin{cases} \beta^{(d)}(N(dt))\,, & \text{in mode 1} \\ -\beta^{(d)}(N(dt))\,, & \text{in mode 2} \end{cases} \tag{9}$$

Our first diffusion limit result, following [25], states that within each mode, the inverse temperature process behaves identically to the case where there is only one mode (i.e. $J = 1$). To state it, we extend the definition of I to

$$I(\beta) \;=\; \begin{cases} \mathrm{Var}_{x\sim f_1^{\beta}}(\log f_1(x))\,, & \beta > 0 \\ \mathrm{Var}_{x\sim f_2^{|\beta|}}(\log f_2(x))\,, & \beta < 0\,. \end{cases} \tag{10}$$

Theorem 1 *Assume the target distribution π is of the form (1), with $J = 2$ modes of weights w_1 and $w_2 = 1 - w_1$, each having iid coordinates as in (3), with tempered distributions as in (2) for an inverse-temperatures list (4) related by (5). Then away from its boundary points 1 and $\beta_{max}^{(d)}$, the process $\{X_t^{(d)}\}$ from (9) converges weakly as $d \to \infty$ to a fixed diffusion process X, which for $X^{(d)} > 0$ satisfies*

$$\begin{aligned} dX_t \;=\; & \left[2\ell^2(X_t)\,\Phi\left(\frac{-\ell(X_t)I^{1/2}(X_t)}{2}\right)\right]^{1/2} dB_t \\ & + \left[\ell(X_t)\,\ell'(X_t)\;\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right. \\ & \left. - \;\ell^2(X_t)\left(\frac{\ell(X_t)I^{1/2}(X_t)}{2}\right)'\phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right]dt\,. \end{aligned} \tag{11}$$

The same equation holds for $X_t < 0$, except with the sign of the drift reversed.

As a check, (11) satisfies the general relation $\mu(x) = \frac{1}{2}\sigma^2(x)\frac{d}{dx}\log\pi(x) + \sigma(x)\sigma'(x)$, which implies that π is *locally invariant* for $X^{(d)}$, i.e. that its generator G has $\pi\,(Gf)(x) = 0$ for appropriate smooth f and for x in the interior of the domain, That is, π is stationary for $X^{(d)}$ locally within each mode, as expected.

However, Theorem 1 describes only what happens on each mode separately; it says nothing about the mode-jumping process itself. Moreover, its state space $(-\infty, -1]\cup[1, \infty)$ is not connected. In fact, we will see below that as $d \to \infty$, the value $\beta^{(d)}_{max}$ will go to infinity and hence never be reached in finite time. To resolve these issues, we make several transformations on the $X^{(d)}_t$ process. First, for $|x| \geq 1$, we define

$$h(x) \;=\; \int_1^{|x|} \frac{1}{\ell(u)} du\,.$$

(For example, in the Exponential Power Family case, $I(\beta) \propto 1/\beta^2$, so (8) gives $\ell(\beta) = I_0^{-1/2}(\beta)\,\ell_0 \propto \beta$, whence $h(x) = \int_1^{|x|} \frac{1}{\ell(u)}\,du \propto \int_1^{|x|} \frac{1}{u}\,du = \log|x|$.) We then set

$$H^{(d)}_{t\,(h(\beta^{(d)}_{max}))^2} \;=\; \mathrm{sign}\left(X^{(d)}_{t\,h(\beta^{(d)}_{max})^2}\right)\;\left[1 + \frac{h\left(X^{(d)}_{t\,h(\beta^{(d)}_{max})^2}\right)}{h(\beta^{(d)}_{max})}\right]\,. \tag{12}$$

Hence, $1 \leq H^{(d)}_t \leq 2$ in the first mode, and $-1 \geq H^{(d)}_t \geq -2$ in the second mode. Also, $H^{(d)}_t$ speeds up $X^{(d)}_t$ by a factor of $h(\beta^{(d)}_{max})^2$, and hence moves at Poisson rate $d\,h(\beta^{(d)}_{max})^2$. This new process $H^{(d)}_t$ satisfies the following.

Theorem 2 *Under the set-up and assumptions of Theorem 1, on $(-2,-1)\cup(1,2)$ (i.e., away from its boundary points), the process $\{H^{(d)}_t\}$ from (12) converges weakly in the Skorokhod topology as $d \to \infty$ to a limiting diffusion H which satisfies*

$$dH_t \;=\; \left[2\,\Phi\left(\frac{-\ell(X_t)I^{1/2}(X_t)}{2}\right)\right]^{1/2} dB_t + \;\ell(X_t)\left[\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right]' dt\,. \tag{13}$$

Furthermore, H leaves constant (uniform) densities locally invariant.

To make further progress, we now use the Proportionality Condition (7), with corresponding inverse-temperature spacing (8). It then follows from the extended definition (10) that $\ell(X_t)\,I^{1/2}(X_t) = \ell_0 r_1^{1/2}$ for $X_t < 0$, and $\ell(X_t)\,I^{1/2}(X_t) = \ell_0 r_2^{1/2}$ for $X_t > 0$, with $\left[\ell(X_t)\,I^{1/2}(X_t)\right]' = 0$ for all $X_t \neq 0$. Hence, Theorem 2 immediately gives:

Corollary 3 *Assume the set-up and assumptions of Theorem 1, and also the Proportionality Condition (7) with inverse-temperature spacing (8). Then as $d \to \infty$, the process $\{H^{(d)}_t\}$ converges weakly in the Skorokhod topology to a limit process H on $(-2,-1)$ and on $(1,2)$, i.e. away from its boundary points. Furthermore, H is a diffusion, with drift 0, and with diffusion coefficient which is constant on each of the two intervals $(-2,-1)$ and $(1,2)$. Specifically,*

$$dH_t \;=\; s(H_t)\;dB_t\,,$$

where $s(H_t) = s_1$ *for* $H_t \in (1,2)$*, and* $s(H_t) = s_2$ *for* $H_t \in (-2,-1)$*, with*

$$s_i \;:=\; \left[2\,\Phi\left(-\frac{1}{2}\ell_0 r_i^{1/2}\right)\right]^{1/2}. \tag{14}$$

Next, we need to join up the two parts of the domain $[-2,-1]\cup[1,2]$ of the process $H_t^{(d)}$. Now, the original process can jump between modes when at the coldest temperture $\beta_{max}^{(d)}$, corresponding to the values ± 2 for the transformed process $H_t^{(d)}$. Hence, we let

$$Z_t^{(d)} \;=\; 2\,\mathrm{sign}(H_t^{(d)}) - H_t^{(d)} \;=\; \begin{cases} 2 - H_t^{(d)}, & H_t^{(d)} \geq 1, \text{ i.e. in mode 1} \\ -2 - H_t^{(d)}, & H_t^{(d)} \leq -1, \text{ i.e. in mode 2} \end{cases} \tag{15}$$

so that $Z_t^{(d)}$ has domain $[-1,1]$ with mode-jumping at 0.

However, by Corollary 3, the limit of the process $Z_t^{(d)}$ will still have diffusion coefficient s_1 or s_2 on its positive and negative parts. We thus rescale the process by setting

$$W_t^{(d)} \;=\; s(Z_t^{(d)})^{-1}\, Z_t^{(d)}. \tag{16}$$

(So, to recap, $W_t^{(d)}$ is defined in (16) in terms of $Z_t^{(d)}$, which is defined in (15) in terms of $H_t^{(d)}$, which is defined in (12) in terms of $X_t^{(d)}$, which is in turn defined in (9) in terms of the original inverse-temperature process $\beta^{(d)}(t)$, which itself arises from running Algorithm 1.) Then $W_t^{(d)}$ has domain $[-\frac{1}{s_2}, \frac{1}{s_1}]$, and limit which is actual Brownian motion on each of $(-\frac{1}{s_2}, 0)$ and $(0, \frac{1}{s_1})$. The precise limit of this process requires the notion of *skew Brownian motion*, a generalisation of usual Brownian motion that, intuitively, behaves just like a Brownian motion except that the sign of each excursion from 0 is chosen using an independent Bernoulli random variable; for further details and constructions and discussion see e.g. [14]. In terms of skew Brownian motion, we have:

Theorem 4 *Under the assumptions of Corollary 3, with* s_i *as in (14), the process* $\{W_t^{(d)}\}$ *from (16) converges weakly in the Skorokhod topology as* $d \to \infty$ *to a limit process* W *which is skew Brownian motion on* $[-\frac{1}{s_2}, \frac{1}{s_1}]$*, with reflecting boundaries, and with excursion probabilities at 0 proportional to* $w_1 s_1$ *(to go positive) and* $w_2 s_2$ *(to go negative).*

The above theorems are all proven in Section 6 below. First, we use them to investigate the computational complexity of the ALPS algorithm.

5 Computational Complexity

Theorem 4 above has implications for the computational complexity of the ALPS algorithm. Indeed, it shows that the limiting process W does not depend at all on the dimension d, and hence has convergence time $O(1)$ as $d \to \infty$. However, W was derived from the processes $H_t^{(d)}$ and $Z_t^{(d)}$, which sped up time by a factor of $(h(\beta_{max}^{(d)}))^2$ from the process

$X_t^{(d)}$, which itself sped up time by a factor d. That is, W was sped up by a total factor of $d[h(\beta_{max}^{(d)})]^2$. So, in the original scaling, the convergence time is $O(d[h(\beta_{max}^{(d)})]^2)$.

More formally, it is shown in [24, Theorem 1] that such diffusion limit convergence implies that for any $\epsilon > 0$, the *convergence time* T_ϵ for each component of the original process to get within ϵ of stationary in Kantorovich-Rubinstein distance, averaged over starting state chosen from stationarity, will be of the same order as the speedup factor. So, combining Theorem 4 and [24, Theorem 1] shows that for the β process of the vanilla ALPS algorithm, the convergence time T_ϵ is $O(d[h(\beta_{max}^{(d)})]^2)$. Furthermore, this convergence time is indeed an appropriate measure of the algorithm's efficiency, since it is proportional to the rate at which the β values can complete a "round trip" from one sample at $\beta = 1$, to a mode jump at $\beta = \beta_{max}$, to another sample at $\beta = 1$, and hence mix well between modes; similar approaches appear in [2, 11, 28].

This raises the question of how $h(\beta_{max}^{(d)})$ grows as a function of d. It is proven in [29] that for the ALPS process to mix modes efficiently, we need the maximum inverse-temperature value $\beta_{max}^{(d)}$ to grow linearly with dimension, i.e. we need to choose $\beta_{max}^{(d)} \propto d$. And, in the Exponential Power Family case, as mentioned above, $I(\beta) \propto 1/\beta^2$ which implies by (8) that $h(x) \propto \log|x|$, so $h(\beta_{max}^{(d)}) \propto \log(d)$. Hence, the complexity order $O(d[h(\beta_{max}^{(d)})]^2)$ equals $O\big(d\,[\log d]^2\big)$. That is, for the inverse temperature process to hit $\beta_{max}^{(d)}$ and hence mix modes takes $O\big(d\,[\log d]^2\big)$ iterations.

If we are not in the Exponential Power Family case, then it may no longer be true that $I(\beta) \propto 1/\beta^2$. However, as $d, \beta \to \infty$, under appropriate smoothness assumptions the densities in the different modes will become approximately Gaussian, which corresponds to the Exponential Power Family case with $r = 2$. And, it is proven in equation (66) of [30] that if the first four moments converge to those of a Gaussian, then $2\beta^2 I(\beta) \to 1$, i.e. approximately $I(\beta) \propto 1/\beta^2$. Hence, from (8), approximately $\ell(\beta) \propto \beta$, so again $h(\beta_{max}^{(d)}) \propto \log(d)$, and the complexity order is still $O\big(d\,[\log d]^2\big)$ as before. We summarise this conclusion as follows.

Corollary 5 *Under the assumptions of Corollary 3, if either* **(a)** *the densities of the two modes of π are in the Exponential Power Family, or* **(b)** *the two modes' first four moments each converge to those of a Gaussian as $d, \beta \to \infty$, then the convergence times T_ϵ for β are $O\big(d\,[\log d]^2\big)$ as $d \to \infty$.*

In a different direction, the paper [30] introduces a *QuanTA Algorithm*, which modifies parallel tempering's usual temperature-swap moves by adjusting the x space in order to permit larger moves in the inverse temperature space. As a result of this, they show [30, Theorem 2] that the resulting $\ell(\beta)$ function is then proportional to $\beta^{k/2}$ for some $k > 2$ (instead of proportional to β). In that case,

$$h(\beta_{max}^{(d)}) \;=\; \int_1^{\beta_{max}^{(d)}} \frac{1}{\ell(u)}\,du \;\le\; \int_1^{\infty} \frac{1}{\ell(u)}\,du \;\propto\; \int_1^{\infty} u^{-k/2}\,du \;=\; (k/2) - 1 \;<\; \infty\,,$$

so that $h(\beta_{max}^{(d)})$ is $O(1)$ rather than $O(\log d)$. This means that the convergence complexity $O(d[h(\beta_{max}^{(d)}]^2)$ becomes simply $O(d)$, i.e. the $[\log d]^2$ factor vanishes. We summarise this observation as follows.

Corollary 6 *Under the assumptions of Corollary 3, if we instead run the version of the ALPS algorithm which uses the QuanTA modification of [30], then the convergence times T_ϵ for β are $O\big(d\,[\log d]^2\big)$ as $d \to \infty$.*

Comparing Corollaries 5 and 6, we see that the QuanTA modification improves the complexity bound by a factor of $[\log d]^2$. This is not surprising, since QuanTA was specifically designed to make the algorithm move faster especially under near-Gaussianity at large β, thus improving the mixing time. This improvement is also borne out through simulation experiments; see [30].

Remark 3 *(More than Two Modes.)* For simplicity, all of the above proofs assumed a mixture of just $J = 2$ modes. However, similar analysis works more generally. Indeed, suppose π is a mixture of $J > 2$ modes, of weights $w_1, w_2, \ldots, w_J \geq 0$ where $\sum_{i=1}^{J} w_i = 1$. Then when $\beta(t)$ reaches $\beta_{max}^{(d)}$, the process chooses one of the J modes with probability w_i (due to the auxiliary mode-jumping step). In this case, a theorem similar to Theorem 4 could be proven by similar methods. The processes $\{W_t^{(d)}\}$ will converge not to skew Brownian motion but to *Walsh's Brownian motion*, a process not on $[-\frac{1}{s_2}, \frac{1}{s_1}]$ but rather on a "star" shape with J different line segments all meeting at the origin (corresponding to $\beta_{max}^{(d)}$). Intuitively, this process behaves as Brownian motion within each segment, but chooses each excursion from the origin using an independent random variable with probabilities w_i; for further details and constructions and discussion see e.g. [3]. (The case $J = 2$ but $w_1 \neq 1/2$ corresponds to skew Brownian motion as in Theorem 4.) This in turn leads to the same complexity bound of $O\big(d\,[\log d]^2\big)$ iterations (or $O(d)$ iterations if using QuanTA) when $J > 2$ as well.

6 Theorem Proofs

In this section, we prove the theorems stated in Section 4 above. Note that Theorem 1 essentially follows directly from previous theoretical analysis of Simulated Tempering in [2, 25], and Theorem 2 then follows from some additional computations using Ito's Formula. By contrast, Theorem 4 requires a different approach, to show that the modified $W^{(d)}$ process converges to reflecting skew Brownian motion, including delicate arguments to show convergence of the corresponding infinitesimal generators especially at the two endpoints and at the excursion point 0.

6.1 Proof of Theorem 1

Since mixing between modes is only possible at $\beta_{max}^{(d)}$, the dynamics for other β will be identical to the single mode case ($J = 1$) as covered in [2, 25]. It therefore follows directly from Theorem 6 of [25] that as $d \to \infty$, the process $\{X_t\}$ converges weakly, at least on $X_t > 0$, to a diffusion limit $\{X_t\}_{t \geq 0}$ satisfying (11). The result for $X_t < 0$ follows similarly.

6.2 Proof of Theorem 2

We assume $x \in (1,2)$; the proof for $x \in (-2,-1)$ is virtually identical. Here $H_t = h(X_t)$, where $h'(x) = \ell(x)^{-1}$, and $h''(x) = -\ell'(x)\ell(x)^{-2}$. Hence, by Ito's Formula,

$$\begin{aligned}
dH_t &= h'(X_t)dX_t + \frac{1}{2}h''(X_t)d\langle X\rangle_t \\
&= \ell(X_t)^{-1}dX_t - \frac{1}{2}\ell'(X_t)\ell(X_t)^{-2}d\langle X\rangle_t \\
&= \ell(X_t)^{-1}\left[2\ell^2(X_t)\Phi\left(\frac{-\ell(X_t)I^{1/2}(X_t)}{2}\right)\right]^{1/2} dB_t \\
&\quad +\ell(X_t)^{-1}\ell(X_t)\ell'(X_t)\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)dt \\
&\quad -\ell^2(X_t)\left(\frac{\ell(X_t)I^{1/2}(X_t)}{2}\right)'\phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)dt \\
&\quad -\frac{1}{2}\ell'(X_t)\ell(X_t)^{-2}2\ell^2(X_t)\Phi\left(\frac{-\ell(X_t)I^{1/2}(X_t)}{2}\right)dt
\end{aligned}$$

In this last equation, the second and fourth terms cancel. Also, since $\Phi' = \phi$, it follows from the chain rule that the third term can be written as

$$-\ell^2(X_t)\left[\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right]' dt\,.$$

This gives (13). Then, writing everything in terms of $H_t = h(X_t)$, this becomes

$$\begin{aligned}
dH_t &= \left[2\,\Phi\left(\frac{-\ell(h^{-1}(H_t))I^{1/2}(h^{-1}(H_t))}{2}\right)\right]^{1/2} dB_t \\
&\quad +\ell(h^{-1}(H_t))\left[\Phi\left(\frac{-I^{1/2}(h^{-1}(H_t))\ell(h^{-1}(H_t))}{2}\right)\right]' dt\,.
\end{aligned}$$

Now, a diffusion of the form $dH_t = \sigma(H_t)dB_t + \mu(H_t)dt$ has locally invariant distribution π provided that $\frac{1}{2}(\log\pi)'\sigma^2 + \sigma\sigma' = \mu$. That holds for constant π if $\sigma\sigma' = \mu$. In this case, we compute that

$$\begin{aligned}
\sigma\sigma' &= \frac{1}{2}(\sigma^2)' = \frac{1}{2}\frac{d}{dH}\left[2\,\Phi\left(\frac{-\ell(h^{-1}(H))I^{1/2}(h^{-1}(H))}{2}\right)\right] \\
&= \frac{1}{2}\left(\frac{dH}{dX}\right)^{-1}\frac{d}{dX}\left[2\,\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right] \\
&= \frac{1}{2}\left(\ell(X)^{-1}\right)^{-1}\left[2\,\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right]' \\
&= \ell(X)\left[\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right]' = \mu\,,
\end{aligned}$$

thus showing that H leaves constant densities locally invariant.

6.3 Proof of Theorem 4

Let $w_{min}^{(d)} = -\frac{1}{s_2}$ and $w_{max}^{(d)} = \frac{1}{s_1}$ be the endpoints of the domain of W. By Corollary 3, $dH_t = s(H_t)\, dB_t$ in the interior of its domain. Since $W_t = s(H_t)^{-1} H_t$, it follows that W_t behaves like Brownian motion on $(-w_{min}^{(d)}, 0)$ and on $(0, w_{max}^{(d)})$. It remains to show that the process converges weakly to skew Brownian motion, including at the boundary points $W_t = 0, w_{min}^{(d)}, w_{max}^{(d)}$. We prove this result using infinitesimal generators, as we now explain.

6.3.1 Method of Proof: Generators

To prove the weak convergence, it suffices by Corollary 8.7 of Chapter 4 of [7] to show (similar to previous proofs of diffusion limits of MCMC algorithms in [21, 22, 4]) that the *infinitesimal generator* $G^{(d)}$ of the process $W^{(d)}$ converges uniformly in x as $d \to \infty$ to the generator G^* of skew Brownian motion, when applied to a *core* $\mathcal{D}$ of functionals, i.e. that

$$\lim_{d\to\infty} \sup_{x\in[w_{min}^{(d)}, w_{max}^{(d)}]} \left| G^{(d)} f(x) - G^* f(x) \right| \;=\; 0\,, \qquad f \in \mathcal{D}\,,$$

where

$$G^{(d)} f(x) \;:=\; \lim_{\delta \searrow 0} \frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta}\,.$$

To this end, let $\mathcal{D}$ be the set of all functions $f : [-w_{min}^{(d)}, w_{max}^{(d)}] \to \mathbb{R}$ which are continuous and twice-continuously-differentiable on $[w_{min}^{(d)}, 0]$ and also on $[0, w_{max}^{(d)}]$, with matching one-sided second derivatives $f''^{+}(0) = f''^{-}(0)$, and skewed one-sided first derivatives satisfying $w_1 s_1 f'^{+}(0) = w_2 s_2 f'^{-}(0)$, and $f'(w_{max}^{(d)}) = f'(w_{min}^{(d)}) = 0$. Then it follows from e.g. [15] and Exercise 1.23 of Chapter VII of [20]) that the generator of skew Brownian motion (with excursion weights proportional to $w_1 s_1$ and $w_2 s_2$ respectively, and with reflections at $w_{min}^{(d)}$ and $w_{max}^{(d)}$) satisfies that $G^* f(x) = \frac{1}{2} f''(x)$ for all $f \in \mathcal{D}$, where $f''(0)$ represents the common value $f''^{+}(0) = f''^{-}(0)$. Furthermore, $\mathcal{D}$ is clearly dense (in the sup norm) in the set of all $C^2[w_{min}^{(d)}, w_{max}^{(d)}]$ functions, so in the language of [7], $\mathcal{D}$ serves as a core of functions for which it suffices to prove that the generators converge.

It follows from Corollary 3, as discussed above, that for any fixed $f \in \mathcal{D}$,

$$\lim_{d\to\infty} \sup_{w\in\left(w_{min}^{(d)}, w_{max}^{(d)}\right)\setminus\{0\}} |G^{(d)} f(w) - G^* f(w)| = 0\,. \tag{17}$$

That is, the generators do converge uniformly to G^*, as required, at least for $w \neq 0, w_{min}^{(d)}, w_{max}^{(d)}$, i.e. avoiding the mode-jumping value 0 and the reflecting boundaries $w_{min}^{(d)}$ and $w_{max}^{(d)}$. To complete the proof, it suffices to prove that (17) also holds at $w = 0, w_{min}^{(d)}, w_{max}^{(d)}$, i.e. to prove

$$\lim_{d\to\infty} G^{(d)} f(0) \;\equiv\; G^* f(0) \;=\; \frac{1}{2} f''(0)\,, \tag{18}$$

$$\lim_{d\to\infty} G^{(d)} f(w_{min}^{(d)}) \;\equiv\; G^* f(w_{min}^{(d)}) \;=\; \frac{1}{2} f''(w_{min}^{(d)})\,, \tag{19}$$

and

$$\lim_{d\to\infty} G^{(d)} f(w_{max}^{(d)}) \;\equiv\; G^* f(w_{max}^{(d)}) \;=\; \frac{1}{2} f''(w_{max}^{(d)})\,. \tag{20}$$

6.3.2 Verification of (19) and (20)

The proofs of (19) and (20) are virtually identical, so here we prove (20).

If the original inverse-temperature process $\beta^{(d)}(t)$ proposes to move in time 1 from inverse-temperature $1+0=1$ to $1+\ell(1)d^{-1/2}$, then by (12), the $H_t^{(d)}$ process proposes to move at Poisson rate $[d\,h(\beta_{max}^{(d)})^2]$ from $1+\frac{0}{h(\beta_{max}^{(d)})}=1$ to

$$1+\frac{h\big(1+\ell(1)d^{-1/2}\big)}{h(\beta_{max}^{(d)})} \;=\; 1+\frac{1}{h(\beta_{max}^{(d)})}\int_1^{1+\ell(1)d^{-1/2}} \frac{1}{\ell(u)}\,du$$

which to first order as $d\to\infty$ is equal to

$$1+\frac{1}{h(\beta_{max}^{(d)})}(\ell(1)d^{-1/2})\frac{1}{\ell(1)} \;=\; 1+\frac{d^{-1/2}}{h(\beta_{max}^{(d)})}\,.$$

Simultaneously, the $Z_t^{(d)}$ process proposes to move from $2-1=1$ to $2-[1+d^{-1/2}/h(\beta_{max}^{(d)})]=1-d^{-1/2}/h(\beta_{max}^{(d)})$, and the $W_t^{(d)}$ process proposes to move from $w_{max}^{(d)}$ to

$$(w_{max}^{(d)})-d^{-1/2}/s_1h(\beta_{max}^{(d)})\,.$$

Let A be the probability that the original $\beta^{(d)}(t)$ process accepts a move from 1 to $1+\ell(1)d^{-1/2}$. Then since $\beta^{(d)}(t)$ proposes to move from 1 to $1+\ell(1)d^{-1/2}$ with probability $1/2$, it actually moves from 1 to $1+\ell(1)d^{-1/2}$ with probability $A/2$, otherwise it stays at 1. So, correspondingly, $W_t^{(d)}$ moves from $w_{max}^{(d)}$ to $(w_{max}^{(d)})-d^{-1/2}/s_1h(\beta_{max}^{(d)})$. Furthermore, recall that $W_t^{(d)}$ moves at Poisson rate $[d\,h(\beta_{max}^{(d)})^2]$, so it moves from $w_{max}^{(d)}$ to $(w_{max}^{(d)})-d^{-1/2}/s_1h(\beta_{max}^{(d)})$ at rate $[d\,h(\beta_{max}^{(d)})^2](A/2)$. However, we instead consider a minor modification of the process $W_t^{(d)}$ which speeds up time by a factor of 2 whenever it is at $w_{max}^{(d)}$, i.e. it moves from there at Poisson rate $[d\,h(\beta_{max}^{(d)})^2](A)$. This is equivalent to the original $\beta^{(d)}(t)$ process "reflecting" by always proposing a positive move from 1, instead of proposing either a positive or a negative (always-rejected) move with probability $1/2$ each. We show in Section 7 below that this minor modification will not change the limiting distribution of the $W_t^{(d)}$, and thus does not affect the proof.

Thus, to first order as $\delta\searrow 0$ [i.e., up to $o(1)$ errors], our modified process $W_t^{(d)}$ will move from $w_{max}^{(d)}$ to $(w_{max}^{(d)})-d^{-1/2}/s_1h(\beta_{max}^{(d)})$ at Poisson rate $[d\,h(\beta_{max}^{(d)})^2](A)$. Hence, setting $x=w_{max}^{(d)}=1/s_1$, we have that

$$\frac{\mathbf{E}[f(W_\delta^{(d)})\,|\,W_0^{(d)}=x]-f(x)}{\delta} \;=\; [d\,h(\beta_{max}^{(d)})^2](A)\Big[f\Big((w_{max}^{(d)})-d^{-1/2}/s_1h(\beta_{max}^{(d)})\Big)-f(x)\Big]+o(1)\,.$$

Then, taking a Taylor series expansion around $x=w_{max}^{(d)}=1/s_1$,

$$\begin{aligned}\frac{\mathbf{E}[f(W_\delta^{(d)})\,|\,W_0^{(d)}=x]-f(x)}{\delta} \;&=\; -[d\,h(\beta_{max}^{(d)})^2](A)[d^{-1/2}/s_1h(\beta_{max}^{(d)})]f'(w_{max}^{(d)})\\ &\quad+\;\frac{1}{2}[d\,h(\beta_{max}^{(d)})^2](A)[d^{-1/2}/s_1h(\beta_{max}^{(d)})]^2f''(w_{max}^{(d)}) \;+\; O(d^{-1/2})+o(1)\\ &=\; -[Ad^{1/2}h(\beta_{max}^{(d)})/s_1]\,f'(w_{max}^{(d)}) \;+\; \frac{1}{2}[A/s_1^2]\,f''(w_{max}^{(d)}) \;+\; O(d^{-1/2})+o(1)\,,\end{aligned}$$

Since $f \in \mathcal{D}$, we have $f'(w_{max}^{(d)}) = 0$, so the first term vanishes. Furthermore, it is shown in [31] that as $d \to \infty$,

$$A \;\to\; 2\,\Phi\left(\frac{-\ell_0}{2\sqrt{r_1}}\right) \;=\; s_1^2\,.$$

Hence,

$$\frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;=\; 0 \;+\; \frac{1}{2}\,[1]\,f''(w_{max}^{(d)}) \;+\; O(d^{-1/2}) + o(1)\,,$$

so that

$$\lim_{d\to\infty} G^{(d)} f(w_{max}^{(d)}) \;=\; \lim_{d\to\infty} \lim_{\delta \searrow 0} \frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;=\; \frac{1}{2} f''(w_{max}^{(d)}) \;=\; G^*(w_{max}^{(d)})\,,$$

as required.

6.3.3 Verification of (18)

To prove (18), note that if the original inverse-temperature process $\beta^{(d)}(t)$ proposes to move in time 1 from $\beta_{max}^{(d)}$ to $\beta_{max}^{(d)} - \ell(\beta_{max}^{(d)})d^{-1/2}$ in one of the two modes (with probabilities w_1 and w_2 respectively), then by (12) the $H_t^{(d)}$ process proposes to move at rate $[d\,h(\beta_{max}^{(d)})^2]$ from $1 + \frac{h(\beta_{max}^{(d)})}{h(\beta_{max}^{(d)})} = 2$ to

$$\pm\left[1 + \frac{h\big(\beta_{max}^{(d)} - \ell(\beta_{max}^{(d)})d^{-1/2}\big)}{h(\beta_{max}^{(d)})}\right] \;=\; \pm\left[2 - \frac{\int_{\beta_{max}^{(d)} - \ell(\beta_{max}^{(d)})d^{-1/2}}^{\beta_{max}^{(d)}} \frac{1}{\ell(u)}\,du}{h(\beta_{max}^{(d)})}\right]$$

$$\approx\; \pm\left[2 - \big(\ell(\beta_{max}^{(d)})d^{-1/2}\big)\,\frac{1}{\ell(\beta_{max}^{(d)})}\right] \;=\; \pm(2 - d^{-1/2})\,.$$

Simultaneously, the $Z_t^{(d)}$ process proposes to move from $2 - 2 = 0$ to $\pm 2 - [\pm(2 - d^{-1/2})] = \pm d^{-1/2}$, and the $W_t^{(d)}$ process proposes to move from 0 to either $d^{-1/2}/s_1$ or $-d^{-1/2}/s_2$. Hence, similar to the above (but without the minor modification), with $x = 0$ we have to first order as $\delta \searrow 0$ that

$$\begin{aligned}&\frac{\mathbf{E}[f(W_\delta) \,|\, W_0 = x] - f(x)}{\delta} \\ &\quad=\; [d\,h(\beta_{max}^{(d)})^2]\left(w_1\alpha_1\left[f\left(d^{-1/2}/s_1\right) - f(0)\right] \;+\; w_2\alpha_2\left[f\left(-d^{-1/2}/s_2\right) - f(0)\right]\right) + o(1)\,,\end{aligned} \tag{21}$$

where α_i is the acceptance probability for the original process to accept a proposal to increase the inverse-temperature from $\beta_{max}^{(d)}$ to $\beta_{max}^{(d)} - \ell(\beta_{max}^{(d)})d^{-1/2}$ in mode i. Now, the argument in [31] shows that as $d \to \infty$ we have

$$\alpha_i \;\to\; 2\,\Phi\left(\frac{-\ell_0}{2\sqrt{r_i}}\right) \;=\; s_i^2\,, \qquad i = 1, 2\,.$$

Hence, taking a Taylor series expansion around $x = 0$, we obtain from (21) that

$$\begin{aligned}
&\frac{\mathbf{E}[f(W_\delta) \,|\, W_0 = x] - f(x)}{\delta} \\
&\quad = \; d\, w_1 s_1^2 \left(d^{-1/2}/s_1\right) f'^{+}(0) + \; \frac{1}{2} d\, w_1 s_1^2 \left(d^{-1/2}/s_1\right)^2 f''^{+}(0) \; + \; O(d\, d^{-3/2}) + o(1) \\
&\quad - \; d\, w_2 s_2^2 \; \left(d^{-1/2}/s_2\right) f'^{-}(0) + \; \frac{1}{2} d\, w_2 s_2^2 \left(d^{-1/2}/s_2\right)^2 f''^{-}(0) + \; O(d\, d^{-3/2}) + o(1) \\
&\quad = \; d^{1/2}[w_1 s_1 f'^{+}(0) - w_2 s_2 f'^{-}(0)] \; + \; \frac{1}{2}[w_1 f''^{+}(0) + w_2 f''^{-}(0)] \; + \; O(d^{-1/2}) + o(1)\,.
\end{aligned}$$

Now, by the definition of $f \in \mathcal{D}$, $w_1 s_1 f'^{+}(0) - w_2 s_2 f'^{-}(0) = 0$, and $w_1 f''^{+}(0) + w_2 f''^{-}(0) = (w_1 + w_2) f''(0) = f''(0)$. Hence, we obtain finally that

$$\frac{\mathbf{E}[f(W_\delta) \,|\, W_0 = x] - f(x)}{\delta} \;\; = \;\; \frac{1}{2} f''(0) + O(d^{-1/2}) + o(1)\,,$$

so that

$$\lim_{d\to\infty} G^{(d)} f(0) \;\; = \;\; \lim_{d\to\infty} \lim_{\delta \searrow 0} \frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;\; = \;\; \frac{1}{2} f''(0) \;\; = \;\; G^*(0)\,.$$

This establishes (18), and hence completes the proof of Theorem 4.

Remark 4 Because of our transformations converting $\beta^{(d)}(t)$ to $W_t^{(d)}$, the limiting process W in Theorem 4 was skew Brownian motion on a continuous interval. It is also possible to consider diffusions directly associated with a discrete graph, see e.g. [8].

7 Appendix: Modified Processes and Occupation Times

Recall that the proof of (20) in Section 6.3.2 above was actually for a minor modification of the process $W_t^{(d)}$, which speeds up time by a factor of 2 whenever it is in the state $w_{max}^{(d)}$. We now argue that this minor modification does not affect the limiting distribution. Indeed, since the modification corresponds to adjusting the rate of time, we can write the modified process as $\widehat{W}_t^{(d)} \equiv W_{\tau_d(t)}^{(d)}$, where $\tau_d(t)$ is the time scale including the occasional speedups. Clearly $\lim_{t\searrow 0} \tau_d(t) = 0$. Also, it follows from Proposition 9 below that the fraction of time that the original process spends at $w_{max}^{(d)}$ converges to 0 as $d \to \infty$. This implies that $\lim_{d\to\infty}(\tau_d(t)/t) = 1$. Since our process $W_t^{(d)}$ is continuous, this means that $\lim_{d\to\infty} |f(W_{\tau_d(t)}^{(d)}) - f(W_t^{(d)})| = 0$. That is, the two processes have the same limiting behaviour as $d \to \infty$. So, the diffusion limit is not affected by making our minor modification as above.

It remains to state and prove Proposition 9. We begin with a result about limiting probabilities for reflecting simple symmetric random walk.

Proposition 7 *Let $\{Y_n\}$ be reflecting simple symmetric random walk on the state space $\{0, 1, 2, \ldots, m\}$, i.e. a discrete-time birth-death Markov chain with transition probabilities $p_{i,i+1} = p_{i,i-1} = 1/2$ for $1 \le i \le m-1$, and $p_{0,1} = p_{m,m-1} = 1$. Then for all $m \in \mathbf{N}$ and all sufficiently large $n \in \mathbf{N}$, $\mathbf{P}(Y_n = 0) \le (2/\sqrt{n}) + (1/m)$. Hence, $\lim_{n,m\to\infty} \mathbf{P}(Y_n = 0) = 0$.*

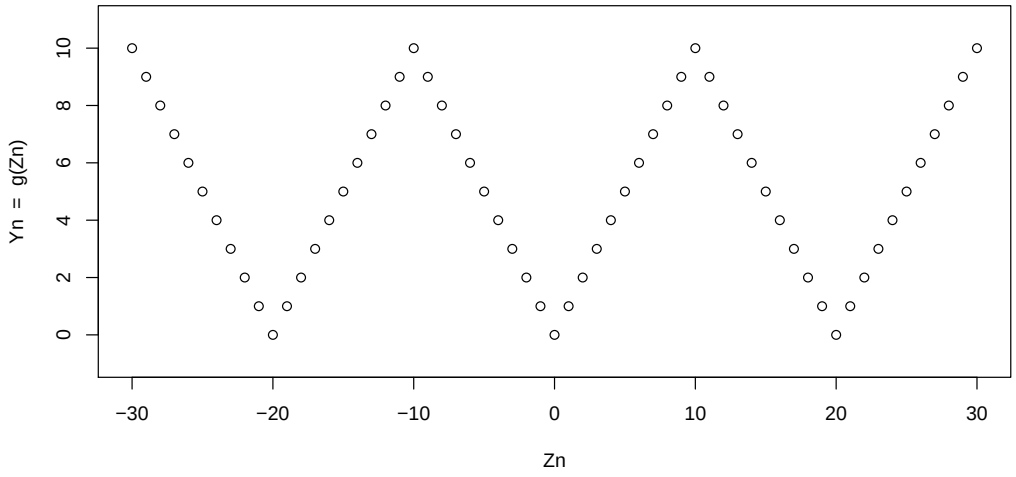

Figure 4: The lifting transformation function "g" (when $m = 10$).

Proof: We condition on $Y_0 = y$; the general case then follows by taking expectation with respect to Y_0. We "lift" $\{Y_n\}$ to $\mathbf{Z}$ by writing $Y_n = g(Z_n)$, where $\{Z_n\}$ is simple symmetric random walk on *all* the integers $\mathbf{Z}$, and $g(z) = \min_j |z - 2jm|$ (see Figure 4). Then

$$\mathbf{P}_y[Y_n = 0] \;=\; \mathbf{P}_y[g(Z_n) = 0] \;=\; \sum_{j \in \mathbf{Z}} \mathbf{P}_y[Z_n = 2jm]$$

$$= \sum_{j \in \mathbf{Z}} \mathbf{P}_y\Big[\text{Binomial}(n, 1/2) = \frac{n}{2} + \frac{y}{2} + jm\Big] \;=\; \sum_{j \in \mathbf{Z}} h\left(\frac{n}{2} + \frac{y}{2} + jm\right) ,$$

where $h(k) = \mathbf{P}[\text{Binomial}(n, 1/2) = k]$. Now, h is maximised when $k = n/2$ (or $(n \pm 1)/2$ if n is odd), and decreases monotonically on either side of that. Hence, find $j_* \in \mathbf{N}$ with $\frac{y}{2} + (j_* - 1)m < 0 \leq \frac{y}{2} + j_* m$. It follows from Stirling's Approximation (see e.g. [27]) that to first order as $n, k, n - k \to \infty$,

$$\mathbf{P}[\text{Binomial}(n, 1/2) = k] \;\leq\; e^{-2n[\frac{1}{2} - \frac{k}{n}]^2} \sqrt{1/2\pi k[1 - (k/n)]}\,,$$

so in particular

$$h\left(\frac{n}{2} + \frac{y}{2} + j_* m\right) \;\leq\; \sqrt{2/\pi n} + o_n(1) \;\leq\; 1/\sqrt{n}$$

for all sufficiently large n, and similarly for $h\left(\frac{n}{2} + \frac{y}{2} + (j_* - 1)m\right)$. Then, by monotonicity, we have for $j > j_*$ that

$$h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\leq\; \frac{1}{m}\Big[h\left(\frac{n}{2} + \frac{y}{2} + (j-1)m + 1\right) + h\left(\frac{n}{2} + \frac{y}{2} + (j-1)m + 2\right) + \ldots + h\left(\frac{n}{2} + \frac{y}{2} + jm\right)\Big] .$$

Hence,

$$\sum_{j>j_*} h\left(\frac{n}{2}+\frac{y}{2}+jm\right) \;\le\; \frac{1}{m}\Big[h\left(\frac{n}{2}+\frac{y}{2}+1\right)+h\left(\frac{n}{2}+\frac{y}{2}+2\right)+h\left(\frac{n}{2}+\frac{y}{2}+3\right)+\dots\Big]\,.$$

But $\sum_k h(k)=1$, so by symmetry $\sum_{k>n/2} h(k)\le 1/2$, and so

$$h\left(\frac{n}{2}+\frac{y}{2}+1\right)+h\left(\frac{n}{2}+\frac{y}{2}+2\right)+h\left(\frac{n}{2}+\frac{y}{2}+3\right)+\dots\le 1/2\,.$$

Thus,

$$\sum_{j>j_*} h\left(\frac{n}{2}+\frac{y}{2}+jm\right) \;\le\; \frac{1}{2m}\,.$$

Similarly,

$$\sum_{j<j_*-1} h\left(\frac{n}{2}+\frac{y}{2}+jm\right) \;\le\; \frac{1}{2m}\,.$$

Therefore, for all sufficiently large n,

$$\sum_{j\in\mathbf{Z}} h\left(\frac{n}{2}+\frac{y}{2}+jm\right) \;\le\; (1/\sqrt{n})+(1/\sqrt{n})+\frac{1}{2m}+\frac{1}{2m} \;=\; (2/\sqrt{n})+(1/m)\,,$$

as claimed. □

Remark 5 Similar arguments show that $\lim_{n,m\to\infty} \mathbf{P}(Y_n=z) = 0$ for any fixed number $z\in\mathbf{N}$, by replacing "$Z_n=2jm$" by "$Z_n=2jm+z$", and "$\frac{n}{2}+\frac{y}{2}$" by "$\frac{n}{2}+\frac{y}{2}-\frac{z}{2}$", throughout the proof, though we do not use that fact here.

Corollary 8 *Let $\{Y_n\}$ be as in Proposition 7. Let $N_0=\#\{i: 0\le i\le n-1,\ Y_i=0\}$ be the occupation time of the state 0 before time n. Then as $n,m\to\infty$, the average occupation time N_0/n converges to 0 in probability.*

Proof: Let $I_i=\mathbf{1}_{Y_i=0}$ be the indicator function of the event $Y_i=0$. Then by Proposition 7, $\lim_{n,m\to\infty}\mathbf{E}[I_n]=\lim_{n,m\to\infty}\mathbf{P}[Y_n=0]=0$. Hence, using the theory of Cesàro sums,

$$\lim_{n,m\to\infty}\mathbf{E}[N_0/n] \;=\; \lim_{n,m\to\infty}\mathbf{E}\Big[\sum_{i=0}^{n-1} I_i\Big]/n \;=\; \lim_{n,m\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{E}[I_i] \;=\; \lim_{n,m\to\infty}\mathbf{E}[I_n] \;=\; 0\,.$$

Hence, by Markov's inequality, since $N_0/n\ge 0$, for any $\epsilon>0$ we have

$$\lim_{n,m\to\infty}\mathbf{P}[(N_0/n)>\epsilon] \;\le\; \lim_{n,m\to\infty}\mathbf{E}[N_0/n]/\epsilon \;=\; 0\,,$$

so that $N_0/n\to 0$ in probability, as claimed. □

Proposition 9 *Let $\{X_n\}$ be a discrete-time birth-death Markov chain on the state space $\{0,1,2,\dots,m\}$, with transition probabilities satisfying that $p_{i,j}=0$ whenever $|j-i|\ge 2$, $p_{i,i+1}=p_{i,i-1}$ for all $1\le i\le m-1$, and $p_{i,i}\le 1-a$ for some fixed constant $a>0$. Let $N_0=\#\{i: 0\le i\le n-1,\ X_i=0\}$. Then as $n,m\to\infty$, N_0/n converges to 0 in probability.*

Proof: Let $\{J_k\}$ be the *jump chain* of $\{X_n\}$, i.e. the Markov chain which copies $\{X_n\}$ except omitting immediate repetitions of the same state, and let $\{M_k\}$ count the number of repetitions. [For example, if the original chain $\{X_n\}$ began $\{X_n\} = (a, b, b, b, a, a, c, c, c, c, d, d, a, \ldots)$, then the jump chain $\{J_k\}$ would begin $\{J_k\} = (a, b, a, c, d, a, \ldots)$, and the corresponding multiplicity list $\{M_k\}$ would begin $\{M_k\} = (1, 3, 2, 4, 2, \ldots)$.] Then the assumptions imply that $\{J_k\}$ has the transition probabilities of reflecting simple symmetric random walk, as in Proposition 7 and Corollary 8 above.

Now, let $K(n)$ be the smallest integer with $M_1 + \ldots + M_{K(n)} \geq n$. Given J_k, the random variable M_k has the Geometric$(1 - p_{J_k J_k})$ distribution, so it is stochastically bounded above by the Geometric(a) distribution, from which it follows that $\lim_{n\to\infty} K(n) = \infty$ w.p. 1. Let $C_s = \#\{i : 0 \leq i \leq K(n),\ J_i = s\}$. Then Corollary 8 implies that $\lim_{n,m\to\infty}(C_0/K(n)) = 0$. On the other hand, N_0 is $\leq$ a sum of C_0 independent Geometric$(1 - p_{00})$ random variables, so $\mathbf{E}[N_0 \,|\, C_0] = C_0/(1 - p_{00}) \leq C_0/a$, and $\mathbf{P}[N_0 > 2C_0/a \,|\, C_0] \to 0$ as $n \to \infty$. Also, $M_1 + \ldots + M_{K(n)-1} \leq n$, and each $M_i \geq 1$, so $n \geq K(n) - 1$. We therefore conclude that

$$\lim_{n,m\to\infty} \frac{N_0}{n} \;\leq\; \lim_{n,m\to\infty} \frac{2C_0/a}{K(n)-1} \;=\; (2/a) \lim_{n,m\to\infty} \frac{C_0}{K(n)} \;=\; 0\,,$$

as claimed. □

Remark 6 It might be possible to instead obtain the conclusion of Proposition 9 via the Ergodic Theorem or the Markov chain Law of Large Numbers, which states that for fixed m the average occupation time N_0/n will converge to the stationary measure of the process at state 0. However, this would require conditions and bounds on the sequence of stationary measures as $m \to \infty$, so it would not be trivial (nor provide the explicit bound of Proposition 7), and we instead use the more direct and quantitative method described herein.

Acknowledgements. We thank Alex Mijatovic and Neal Madras for very helpful comments related to Section 7 herein, and thank the editor and referees for very insightful suggestions which greatly improved the manuscript. This research was partially supported by EPSRC grants EP/R018561/1 and EP/R034710/1 to GOR, and NSERC grant RGPIN-2019-04142 to JSR.

References

[1] Emile Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines*. New York, NY; John Wiley and Sons Inc., 1988.

[2] Yves F Atchadé, Gareth O. Roberts, and Jeffrey S. Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, 2011.

[3] Martin T. Barlow, Jim Pitman, and Marc Yor. On Walsh's Brownian motions. *Séminaire de probabilités (Strasbourg)*, 23:275–293, 1989.

[4] Myléne Bédard and Jeffrey S. Rosenthal. Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *The Canadian Journal of Statistics*, 36:483–503, 2008.

[5] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

[6] S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng (eds). *Handbook of Markov chain Monte Carlo*. Chapman & Hall, 2011.

[7] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

[8] Mark Freidlin and Matthias Weber. On random perturbations of hamiltonian systems with many degrees of freedom. *Stochastic Processes and their Applications*, 94:199–239, 2001.

[9] Charles J Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics*, 23:156–163, 1991.

[10] W. Keith Hastings. Monte Carlo Sampling Methods Using Markov chains and their Applications. *Biometrika*, 57(1):97–109, 1970.

[11] Saul Jacka and Ma. Elena Hernández-Hernández. Minimising the expected commute time. *Stochastic Processes and their Applications, to appear*, 2019.

[12] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[13] Aminata Kone and David A Kofke. Selection of Temperature Intervals for Parallel-Tempering Simulations. *The Journal of Chemical Physics*, 122(20):206101, 2005.

[14] Antoine Lejay. On the constructions of the skew Brownian motion. *Probability Surveys*, 3:413–466, 2006.

[15] Thomas M. Liggett. *Continuous Time Markov Processes: An Introduction*. American Mathematical Society, 2010.

[16] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451, 1992.

[17] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[18] M. Pincus. A Monte-Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems. *Journal of the Operations Research Society of America*, 18(6):967–1235, 1970.

[19] Cristian Predescu, Mihaela Predescu, and Cristian V Ciobanu. The Incomplete Beta Function Law for Parallel Tempering Sampling of Classical Canonical Systems. *The Journal of Chemical Physics*, 120(9):4119–4128, 2004.

[20] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Springer, 3rd edition, 2004.

[21] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

[22] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[23] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[24] Gareth O. Roberts and Jeffrey S. Rosenthal. Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *The Journal of Applied Probability*, 24(53):1–11, 2014.

[25] Gareth O. Roberts and Jeffrey S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *The Annals of Applied Probability*, 24(1):131–149, 2014.

[26] Jeffrey S. Rosenthal. Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7(13):123–128, 2002.

[27] Jeffrey S. Rosenthal. Maximum Binomial Probabilities and Game Theory Voter Models. *Advances and Applications in Statistics, to appear*, 2020.

[28] Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *arXiv preprint arXiv:1905.02939*, 2020.

[29] Nicholas G. Tawn, Sigurd Assing, Matt Moores, and Gareth O. Roberts. The Annealed Leap-Point Sampler (ALPS) for Multimodal Target Distributions. *In preparation*, 2021.

[30] Nicholas G. Tawn and Gareth O. Roberts. Accelerating Parallel Tempering: Quantile Tempering Algorithm (QuanTA). *Applied Probability Trust, to appear*, 2018.

[31] Nicholas G. Tawn, Gareth O. Roberts, and Jeffrey S. Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, 30:27–41, 2020.

[32] Dawn B Woodard, Scott C Schmidler, and Mark Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804, 2009.